Performance study of DTW-based spike measurement anomaly detection in sensors on real-world tests

Bram Cornelis, Federico Deuschle, Konstantinos Gryllias

Performance study of DTW-based spike measurement anomaly detection in sensors on real-world tests

Bram CORNELIS[1], Federico DEUSCHLE[1,2], Konstantinos GRYLLIAS[2,3]

[1]Siemens Digital Industries Software NV, Belgium
bram.cornelis@siemens.com
[2]KU Leuven, Department of Mechanical Engineering, Belgium
[3]Flanders Make @ KU Leuven, Belgium

Abstract

Automated anomaly detection in sensor data plays a crucial role in various applications, including predictive maintenance, quality control, and prototype testing in the automotive industry. This paper focuses on a specific type of anomaly, known as "Spikes", which are sharp, sudden outlier values with no correlation to surrounding samples. The study presents a novel Dynamic Time Warping (DTW) based technique for detecting these spikes in online, multi-channel acquisitions during automotive testing. The technique has been validated on a real-world dataset acquired during a measurement campaign on an electric vehicle. The dataset consists of both anomalous and non-anomalous signals with varying dynamic ranges, patterns, lengths, and sensor types. The results show the method's accuracy in avoiding false positives, such as mistaking spikes for other physical impulses during the test, like tires squeaking, or any other physical impulse coming from the engine or other sub-component of the vehicle under test.

1 Introduction

Detection of anomalies in timeseries data is an important task in many industries [1]. In the automotive sector, sensor data is collected in various stages of the vehicle lifecycle [2]. During the development phase extensive measurement campaigns are conducted on physical prototypes. After manufacturing and assembly quality control tests are performed at the end of the production line. Finally, when the vehicles are in operation at the end users, sensor data can be used to monitor the vehicle fleet, for example enabling predictive maintenance.

In this work we focus on the physical prototype testing which happens during the vehicle development phase. The vehicle is traditionally heavily instrumented with different types of physical sensors that are needed to characterize the dynamical behaviour of a newly designed variant of a certain vehicle. This is a time-intensive and error-prone activity as various problems can occur within the measurement chain. Most common hardware problems that can be found are misconnection issues due to defective cables or sensors, ADC stuck values, EMI due to other sources functioning close to the sensors, ground loop noise, etc. These problems can be observed in the raw measured time series as spikes, drifts, noise, or stuck-at-constant anomalies [3].

While various time series anomaly detection methods have been proposed in literature (cf. survey in [1]), several of them have prohibitive computational complexity or memory requirements, while performance can vary over different datasets and anomaly types. In this work, we focus on a specific sensors anomaly type which commonly occurs in Noise, Vibration and Harshness (NVH) and durability testing applications, i.e., "spikes". Spikes appear in the measured time series as unexpected, sharp peaks with no correlation to previous samples. However, in these testing applications there may also be physical impulses or shock events (non-anomalous patterns) which can be confused with actual spike sensor faults (anomalous patterns). The distinction between the two cases can only be made by an experienced test engineer.

As the state-of-the-art time series anomaly detectors were not designed for this specific NVH testing application, they may not be able to properly distinguish between the actual spike anomalies and the physical impulses and shock events (which are not anomalous). Therefore, the authors previously proposed [4] a novel spike detection approach based on Dynamic Time Warping (DTW), which is designed to make this distinction. In the current work, the method will be further validated on a real-world dataset, which was acquired during a measurement campaign on an electric vehicle. The paper is organized as follows. Section 2 presents the methodological foundations of the proposed approach. Section 3 discusses the real-world dataset and measurement campaign which was conducted for this study. Section 4 presents the results and discussion. Finally, Section 5 reports the conclusions of the work.

2 Methodology

2.1 Problem statement

The goal of this work is to build an automatic sensor spike detection method which is capable of distinguishing between physical peaks (PP) and artificial spikes (AS). The PP are the natural impulses and shock events which are observed in structural response measurements. These have a physical root cause (e.g., a vehicle driving over a pothole) and should not be flagged as anomalous events. The AS on the other hand are caused by faults in the measurement chain, e.g., sensor malfunction, worn cables, etc. These AS should be flagged as anomalous events, such that the test operator is automatically notified of the issue and can take corrective actions. We assume that the difference between PP and AS is observable in the waveform shapes as follows:

• **Physical Peaks (PP)** show a slow oscillating decay after the maximum value is reached. The decay time and oscillation frequency are determined by the mechanical system properties (mass, damping, stiffness).

• **Artificial Spikes (AS)** can be considered as outliers with no dynamic response (e.g., no correlation with previous and subsequent samples).

To illustrate these concepts, Figure 1 shows signals from real world test data. The difference between an AS (on the left) and a PP (on the right) can clearly be observed.

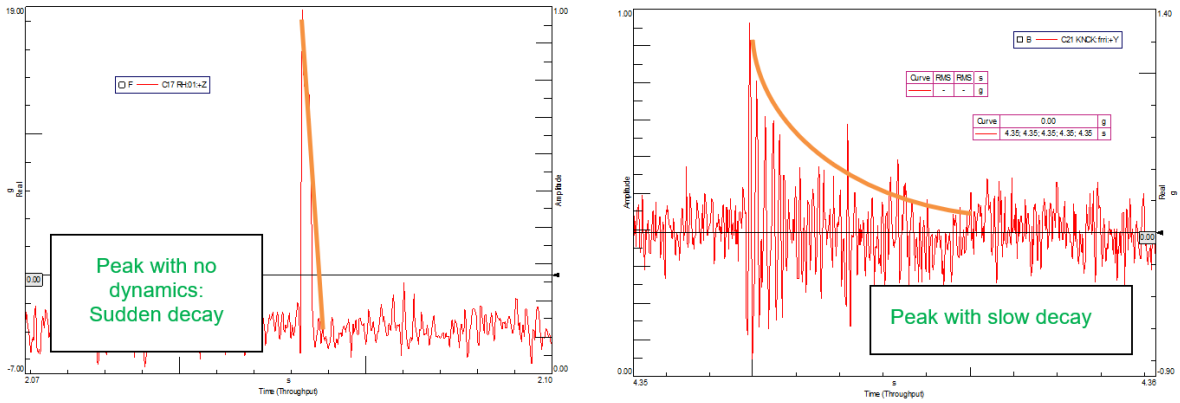

Figure 1: Real-world examples of artificial spikes (left) and physical peaks (right).

2.2 Spike detection methodology

In [4], a four-step spike detection methodology was first introduced, cf. Figure 2. This methodology first extracts a number of candidate peaks (steps 1 to 3) from the raw measurement signal. Then, in a final step, each candidate peak is evaluated by a Dynamic Time Warping (DTW) method, which evaluates the waveform shape of the candidate peak and classifies it as either a PP or AS. The following subsections will further elaborate the four steps of the spike detection method.

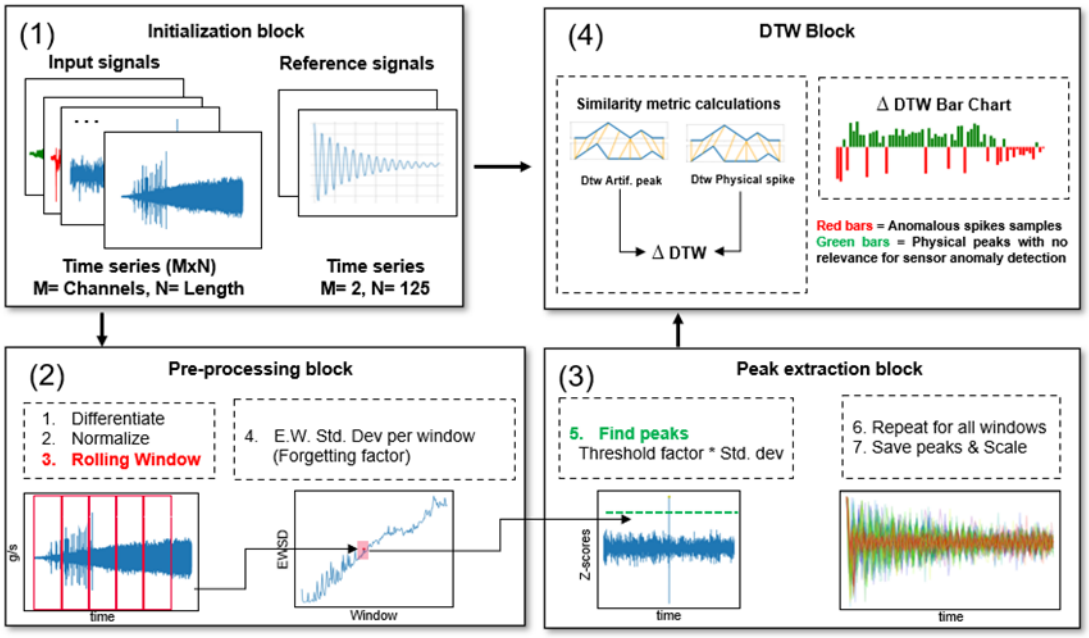

Figure 2: DTW-based spike detector - processing blocks.

2.2.1 Definition of reference signals

To evaluate whether a potential peak resembles more an AS or a PP, idealized reference signals for both cases have to be constructed. For the PP, it is assumed that the signal will exhibit a slow oscillating decay with the decay time and oscillation frequency determined by the mechanical system properties (mass, damping, stiffness). For a Single-Degree-Of-Freedom (SDOF) linear system which is excited by an impulsive force, the dynamic behavior can be expressed as:

$$mX'' + cX' + kX = 0, \tag{1}$$

where m, c and k are the mass, damping and stiffness of the structure and X'', X' and X are the acceleration, velocity and displacement. The response of more complex real-world structures corresponds to a sum of several independent frequencies and damping ratios which leads to more complex waveform shapes [5]. Nevertheless, for the case of artificial spikes, such dynamic behaviour cannot be observed. Figure 3 shows the reference template signals of AS and PP that are used in this work.

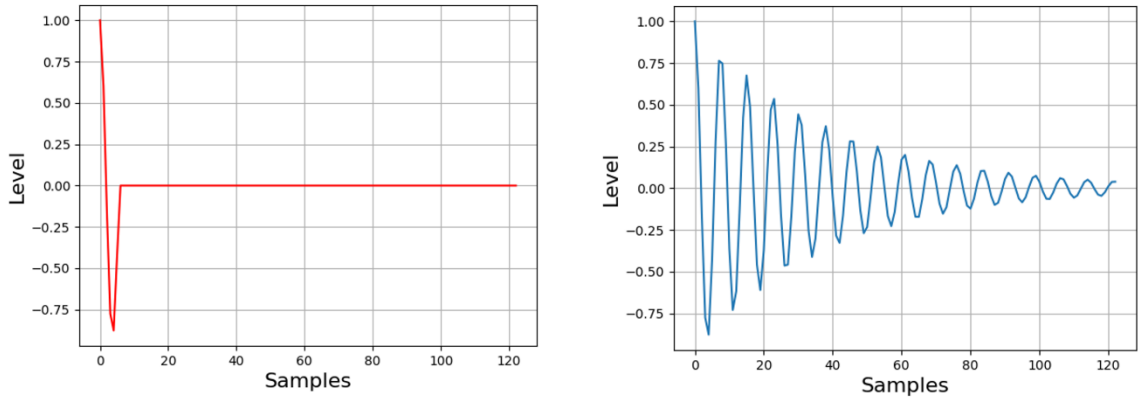

Figure 3: Reference templates of artificial spike (left) and physical peak (right).

2.2.2 Pre-processing

We consider a sensor time series that will be checked as an isolated signal, referred to as $x(t)$. The algorithm starts by differentiating the signal to enhance the presence of the most relevant changes in the gradient $x'(t)$. Then, Z-scores are calculated to normalize the time series, this means to subtract each sample by the mean value and divide by the overall standard deviation [6]. We will define this output variable as $z(t)$. After this first processing, the signal is divided into equally sized window slices. For this case, the window size is 2048 samples and for each window slice an exponentially weighted standard deviation (EWSD) calculation is performed with a forgetting factor $\alpha = 0.3$. The EWSD is calculated in the same way as the exponentially weighted moving average (EWMA) [7], i.e.:

$$\mathrm{EWSD}_k = \alpha * \hat{\sigma}_k(z) + (1-\alpha) * \mathrm{EWSD}_{k-1} \tag{2}$$

where k is the window slice index, α is a forgetting factor between 0 and 1, EWSD_{k-1} corresponds to the EWSD from the previous window slice and $\hat{\sigma}_k(z)$ corresponds to the standard deviation of $z(t)$ in window slice k.

2.2.3 Peak extraction

Following the previously described pre-processing procedure, an EWSD value per window slice is obtained. In order to identify potential peaks, a dynamic threshold is then set as EWSD*Factor. The value of Factor will tune the sensitivity of the algorithm to find more or less peaks to be further analysed by the DTW method. Each peak that surpasses the threshold and the successive 125 samples are extracted and stored for further analysis. 125 samples correspond to a timeframe close to 2.5 ms for the sampling frequency used in these measurements (i.e., 51200 Hz). The extracted peaks are scaled to unit amplitude to reduce the influence of amplitude while comparing with the reference templates (cf. next section).

2.2.4 Dynamic Time Warping

Once the peaks have been extracted from the pre-processed time series, each one of them is compared to the reference templates (cf. Figure 3). Since the peaks were all extracted from their maximum value and then scaled with this same value, it means that they all start at amplitude equal to 1. Nevertheless, since the frequencies and damping ratios will vary from each specimen under test, the use of Dynamic Time Warping (DTW) is essential to perform waveform shapes studies. DTW, which was introduced in speech recognition applications [8], allows to measure the similarity between two temporal sequences which may vary in speed, in contrast to the traditional Euclidean distance. DTW finds the optimal alignment between two time series and captures flexible similarities by aligning the coordinates inside both sequences. Figure 4 shows graphically how the most similar elements are linked by a grey solid line. The cost of the optimal alignment can be recursively computed by the following expression:

$$D(i,j) = \zeta(x_i, y_j) + \min\begin{Bmatrix} D(i-1,j-1) \\ D(i,j-1) \\ D(i-1,j) \end{Bmatrix}, \quad i = 1 \dots N, \; j = 1 \dots M, \tag{3}$$

where $x = (x_1, \dots, x_N)$ and $y = (y_1, \dots, y_M)$ are two time series and where ζ is the distance between 2 elements. These terms form D, a N-by-M matrix that will allow to establish the "least-costly" warping path. Intuitively, the distance function has a small value when the sequences are similar and a large one if they are different. Thus, DTW finds the optimal path that runs through the low-cost "valleys" in the cost matrix. For this research, all DTW related calculations were executed with the DTAIDistance Python package [9].

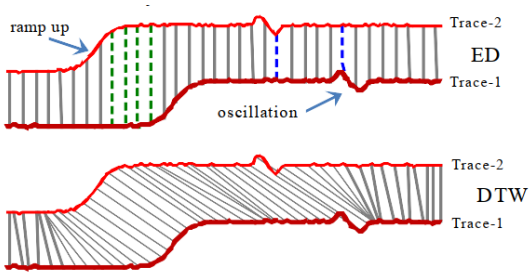

Figure 4: Comparison between Euclidean Distance (ED) and Dynamic Time Warping (DTW). Grey lines indicate linked elements between time series [10].

Using DTW, two similarity metrics are now calculated between each extracted candidate peak and the AS and PP reference template signals. The lowest between both will define if the peak is considered as anomalous or not (e.g., if the lowest similarity metric corresponds to the comparison against the AS template, then the peak will be considered as anomalous). The subtraction of these two metrics leads to the creation of the ΔDTW for every peak:

$$\Delta DTW = DTW_{AS} - DTW_{PP} \tag{4}$$

$$\begin{cases} If\ \Delta DTW > 0 \ \Rightarrow\ Not\ Anomaly \\ If\ \Delta DTW < 0 \ \Rightarrow\ \ Anomaly \end{cases} \tag{5}$$

Since traditional implementations of DTW have a computational complexity of $O(N^2)$ [11], where N is the number of samples to be fed into the DTW block, we keep N small by only applying DTW to the suspicious peaks and successive 125 samples which were obtained in the peak extraction block. This procedure leads to short computation times (as $N = 126$), which allow for online use while a test is conducted.

3 Real-world dataset: vehicle measurement campaign

3.1 Measurement campaign setup

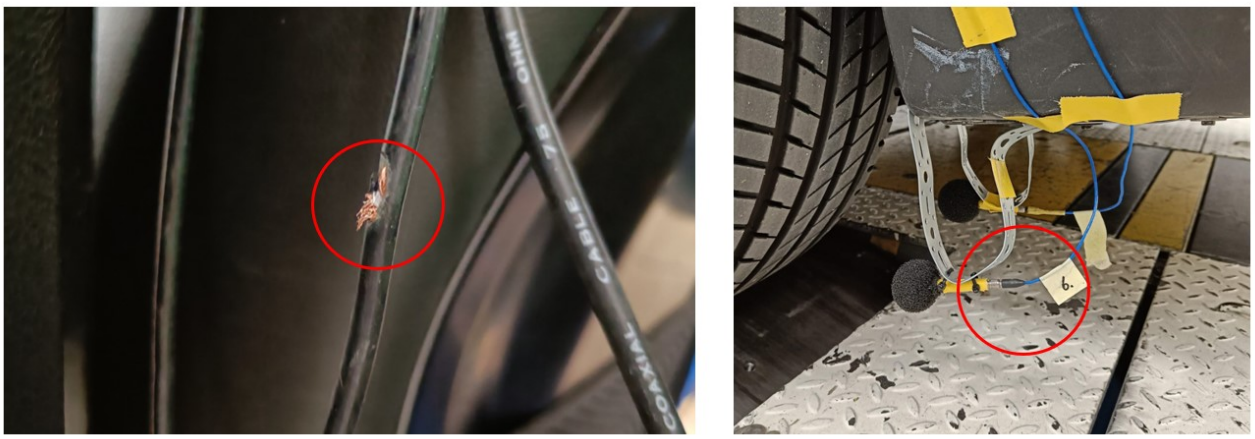

Figure 5: Examples of damaged cable (left) and loosened screw connection (right).

To create a real-world benchmark dataset to validate the spike detector methodology, a test campaign was conducted on an electric vehicle driving on a chassis dynamometer. During these tests sensor anomalies were injected on purpose by using several damaged cables for specific sensors mounted on the vehicle (accelerometers, tachometer and microphones), while for others, the connection screw on the sensor side was loosened to create short-duration signal cuts, cf. Figure 5. For both cases, small and large amplitude spikes were indeed observed in the measured time series. The vehicle was driven in run-up and run-down sessions as

is typical in Noise, Vibration and Harshness (NVH) measurements, hence creating non-stationary test conditions. After each measurement run, the test operator performed a visual inspection of the time series for each of the known "faulty" sensors, to verify whether spike anomalies could be observed.

3.2 Spike detector validation dataset

While during the test campaign other types of sensor anomalies besides spikes were observed (e.g., signal dropouts, high noise levels, flatlining, …), in this work we only focus on the measurement runs and measurement channels where spike anomalies were observed. A set of 24 time series were selected to evaluate the performance. They are a combination of 12 anomalous signals (time series with verified AS), and another set of 12 healthy signals without any visible or audible AS events.

Examples of a healthy signal and an anomalous signal are illustrated in Figure 6 and Figure 7 respectively. It can be observed that in the healthy signal, sometimes a large peak occurs. Upon closer inspection, the peak however shows decaying oscillations indicative of structural dynamic behaviour (cf. Figure 6, zoomed-in view). Hence this corresponds to a PP which should not be indicated as an anomaly.

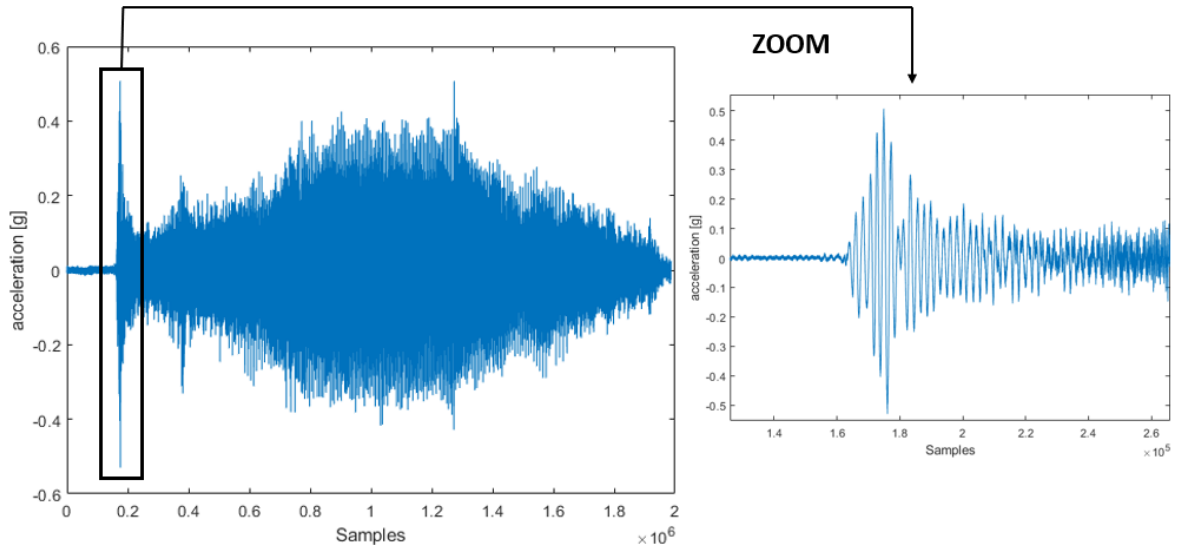

Figure 6: Example of healthy vibration signal with a PP.

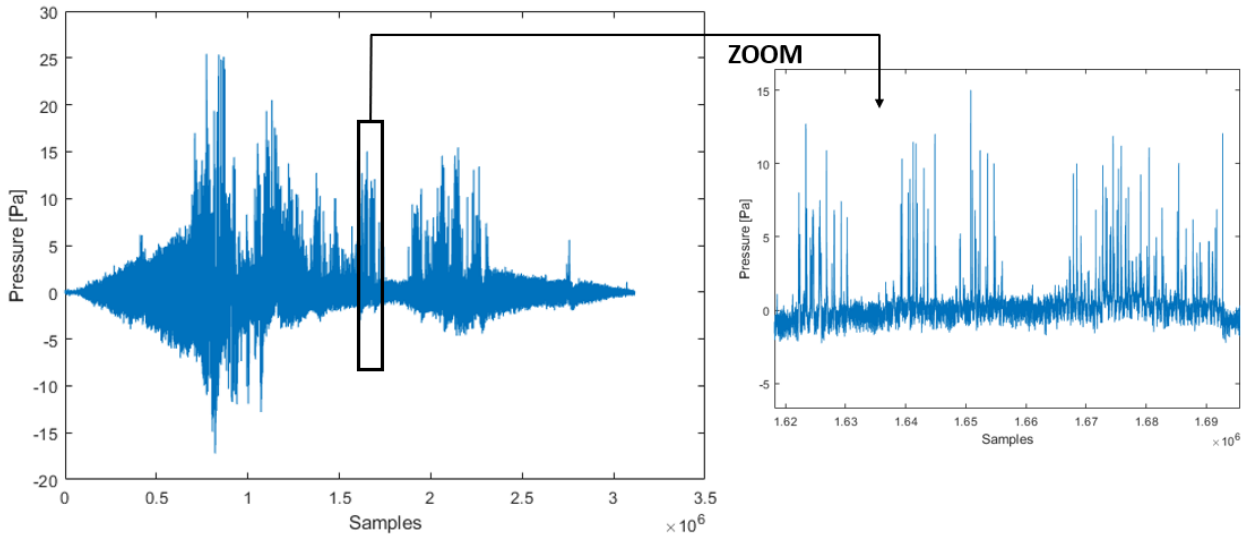

Figure 7: Example of an anomalous microphone signal with multiple AS.

4 Results and discussion

4.1 Ground truth labelling by reference algorithm

While the selected anomalous signals were known to contain multiple AS, it would require significant effort to label each individual AS, while moreover the labelling quality may be compromised due to being a subjective activity. Hence for validation purpose, labels were instead provided by a reference algorithm. In this work the Bayesian spike detector algorithm of [12] was selected for this purpose. This algorithm was validated in previous work and was known to give acceptable performance for NVH measurement applications, but has as disadvantage that its computation time can be high for time series which are affected by many AS. Moreover, the method does not explicitly try to distinguish between AS and PP, so that it may erroneously classify a PP as an AS (i.e., a false positive).

The Bayesian spike detector of [12] aims to precisely detect all samples which are affected by an AS, as the method also aims at correcting the affected samples and replacing them by a prediction of the underlying (healthy) signal. The output of the Bayesian detector is therefore a Boolean indicator function with the same length as the original time series (`0´corresponding to a sample without AS and `1´ corresponding to a sample affected by an AS). The DTW-based spike detector of this work however only aims at detecting each individual AS and does not aim at correcting the signal. It therefore only gives an indication of occurrence of an AS, but does not output a precise indicator function which indicates each sample that is affected by an AS. In particular, in most cases the DTW-based spike detector only indicates a single data point within an AS. Moreover, it (by design) disregards peaks which are in close proximity of an already extracted peak during the peak extraction step.

To be able to compare the outputs of the two spike detection methods, an alignment step is applied whereby the outputs are compared within a prescribed observation window (e.g., 20 ms). If the indicator functions both have one or more entries of `1´ within this observation, it is considered that they both detected the same AS.

4.2 Performance metrics

After the alignment step explained in previous section, we consider a True Positive (TP) to be a case where both the DTW-based spike detector and the reference algorithm indicate the presence of an AS, a False Positive (FP) where only the DTW-based spike detector indicates the presence of an AS, and a False Negative (FN) where only the reference algorithm indicates the presence of an AS. Based on these definitions, we calculate the Precision and Recall as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{7}$$

A high precision value (close to 1) implies that only few false positives are found, which is considered an important property (otherwise, the test operator would too often stop the tests for no reason). A high recall value on the other hand implies that every true AS is indeed detected. While this is ideally also as high as possible (close to 1), we consider that in this application a small number of missed detections can be tolerated from the user perspective.

4.3 Results for anomalous signals

Figure 8 displays the Precision and Recall results averaged over the 12 selected anomalous signals. The perfect spike detector would achieve Precision and Recall scores of 1, and thus be situated in the upper right corner of the graph.

As indicated in Section 4.1, an observation window has to be selected within which the DTW-based spike detector output and reference algorithm outputs have to be aligned. As this choice has a direct impact on the

obtained Precision and Recall scores, the results are shown for 3 different choices (indicated by green, red and purple coloured lines).

Another important parameter is the threshold Factor which is used in the peak extraction step of the DTW-based spike detector. The higher this threshold, the less candidate peaks are passed into the DTW step, such that there is less chance of erroneously classifying a peak to be an AS. Therefore, higher values lead to higher Precision (i.e., less false alarms), but at the same time also to lower recall scores (more missed detections).

The results show that even for the smallest observation window, a high Precision score (above 0.8) can be achieved. The Precision goes up to 0.95 for the larger observation window of 51200 samples (=1 second). For this setting, an acceptable Recall above 0.8 can also be achieved at the same time. As indicated before, we consider that a high Precision score is important in this application, while lower values for Recall can still be acceptable from the user perspective.

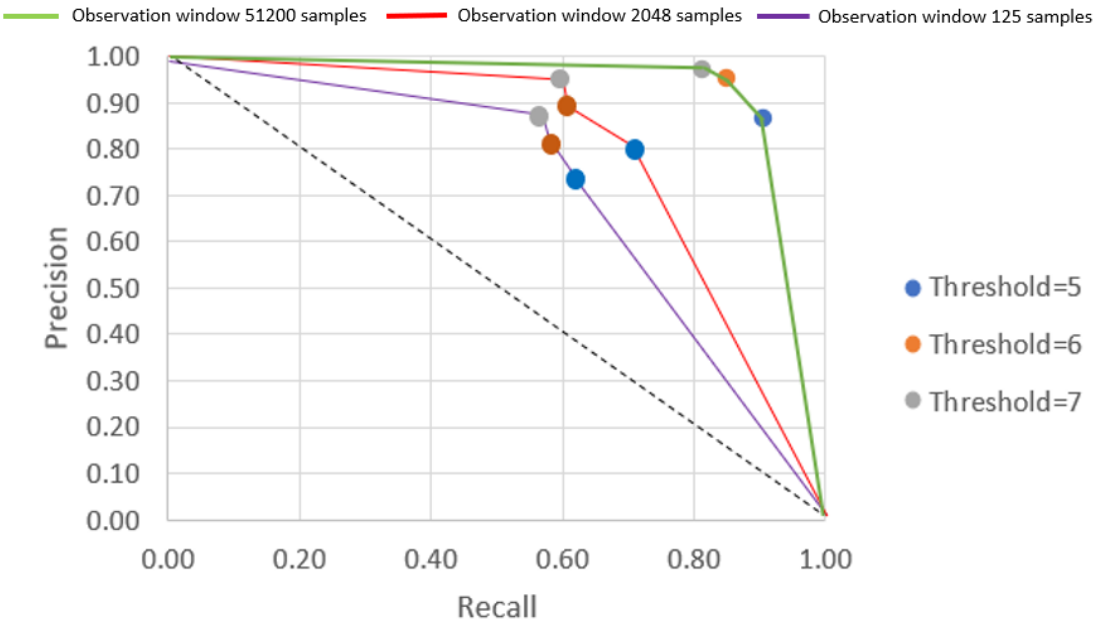

Figure 8: Average Precision and Recall scores for anomalous signals.

Another important aspect is the required computation time of the methods. Figure 9 reports the computation times which were required to process each of the 12 anomalous signals. Each timeseries had a duration of 40-50 seconds, corresponding to more than 2 million data points at a sampling rate of 51200 Hz. It can be observed that the Bayes reference algorithm requires long computation times in some cases, in

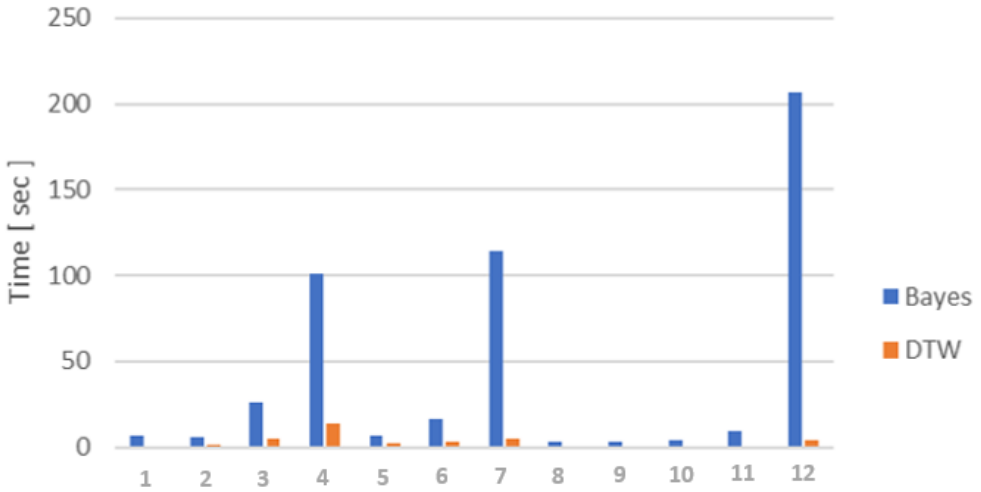

Figure 9: Computation time of Bayes reference algorithm and DTW-based algorithm, for 12 anomalous signals.

particular when many AS occur in the signal. The computation times of the DTW-based method are always significantly shorter and are considered to be acceptable for this application.

4.4 Results for healthy signals

As a final sanity check, we also report the number of AS which are found by the algorithms for the 12 selected "healthy" signals. These were signals without any visible or audible AS events, hence where the ground truth is assumed to be that there are no AS. Figure 10 shows the amount of AS which were erroneously found for each of the 12 healthy signals, both by the Bayes reference algorithm and the DTW based algorithm. It can be observed that the Bayes algorithm finds more AS than the DTW-based algorithm, except for 1 case which requires further investigation. A possible explanation for this is that the Bayes algorithm sometimes confuses a PP for an AS (e.g., in cases like the one illustrated in Figure 7), while the DTW algorithm correctly avoids that the PP leads to a false alarm.

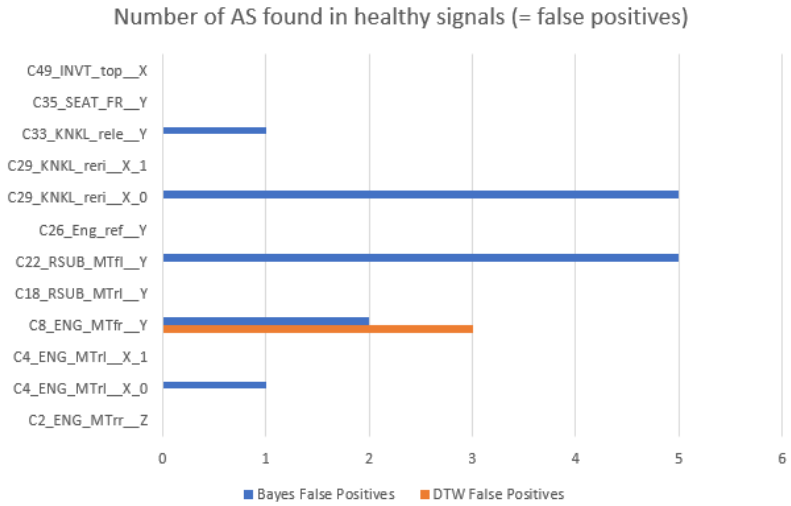

Figure 10: Amount of AS found in 12 selected healthy signals by Bayes reference algorithm and DTW-based algorithm.

5 Conclusions

This paper presented a spike detection method based on DTW with low computational cost, which is suitable for detecting sensor spike anomalies during measurement campaigns on physical prototypes in the vehicle development phase. The method consists of 4 processing blocks which pre-process the sensor signal and extract candidate peaks, which are then compared to reference template signals through DTW in order to classify them as either physical peaks (non-anomalous) or artificial spikes (anomalous). The method was validated on a real-world benchmark dataset, which was measured during a test campaign that was conducted on an electric vehicle driving on a chassis dynamometer. It was demonstrated that the DTW-based method achieves high Precision and Recall scores, at a significantly lower computation time compared to a reference algorithm. Moreover, it raises less false alarms, which can be explained by the fact that it is designed to distinguish physical peaks in the data from true anomalous artificial spikes.

Acknowledgments

We gratefully acknowledge the European Commission for its support of the Marie Sklodowska Curie program through the H2020 ETN MOIRA project (GA 955681).

References

[1] S. Schmidl, P. Wenig, and T. Papenbrock, “Anomaly detection in time series: a comprehensive evaluation”, *Proceedings of the VLDB Endowment*, 2022, 15(9), pp.1779-1797.

[2] C. Derse, M. El Baghdadi, O. Hegazy, U. Sensoz, H.N. Gezer, and M. Nil, “An Anomaly Detection Study on Automotive Sensor Data Time Series for Vehicle Applications” *Proceedings Sixteenth International Conference on Ecological Vehicles and Renewable Energies (EVER),* 2021 May 5, pp. 1-5, IEEE.

[3] K. Ni, N. Ramanathan, M. N. H. Chehade, L. Balzano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen and M. Srivastava, “Sensor Network Data Fault Types”, *ACM*, 2009, vol. 5, no. 3, pp. 1-29.

[4] F. Deuschle, B. Cornelis, and K. Gryllias, “Robust sensor spike detection method based on Dynamic Time Warping”, in *Proceedings International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2022)*, Corfu, Greece, 2022 October 19-21, pp. 20-27.

[5] C. Harris, “Shock and vibration handbook”, McGraw-Hill, 2002.

[6] F. Deuschle, B. Cornelis, J. Lanslots and K. Gryllias, “Overload detection in MEMS microphones-based acoustic arrays”, in *Proceedings ISMA Conference*, Leuven, Belgium, 2022.

[7] T. Fehlmann and E. Kranich, “Exponentially Weighted Moving Average (EWMA) prediction in the Software Development Process”, in *Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement*, Rotterdam, Netherlands, 2014.

[8] F. Petitjean, A. Ketterlin and P. Gançarski, “A global averaging method for dynamic time warping with applications to clustering”, *Pattern Recognition*, 2011, vol. 44, pp. 678-693.

[9] W. Meert, K. Hendrickx, T. Van Craenendonck and P. Robberechts, “DTAIDistance (Version 2) [Computer software]”, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3981067.

[10] Y. Chen, B. Hu, E. Keogh and G. E. Batista, “DTW-D: Time Series Semi-Supervised Learning from a Single Example”, in *Proc. ACM SIGKDD*, Chicago, USA, 2013.

[11] A. Mueen and E. Keogh, “Extracting Optimal performance from Dynamic Time Warping”, in *Proc. ACM SIGKDD*, San Francisco, USA, 2016.

[12] B. Cornelis and B. Peeters, “Online Bayesian spike removal algorithms for structural health monitoring of vehicle components”. In *Proceedings of the 9th International Conference on Structural Dynamics (Eurodyn)*, 2014, Porto, Portugal, pp. 2295–2301.